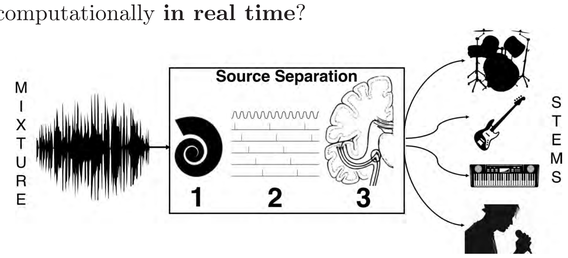
Versatile Scaling for Neurally Inspired Auditory Source Separation

Haydn Jones
New Mexico Institute of Mining and Technology
haydn.jones@student.nmt.edu

Mohit Dubey
Oberlin College and Conservatory
mdubey@oberlin.edu

Background

Separating multiple instruments from a song is an unsolved problem in signal processing. However, this problem is solved instantaneously by the human auditory system. Can we mimic what the brain is doing to solve this problem computationally **in real time**?



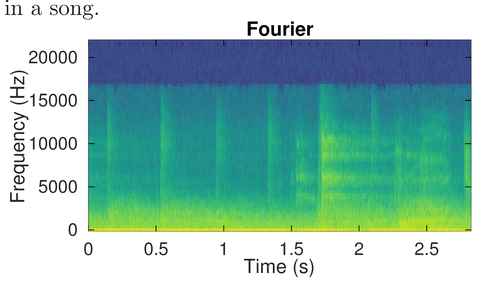
The human auditory pathway uses three primary principles to encode sound:

? 1. **Spectral Representation**: the cochlea produces a time-frequency representation that is logarithmic in frequency and amplitude
✓ 2. **Phase Preservation**: groups of ascending auditory neurons fire in phase with incoming sound waves
✓ 3. **Sparse Coding**: auditory cortex is a highly overcomplete representation of the cochlea

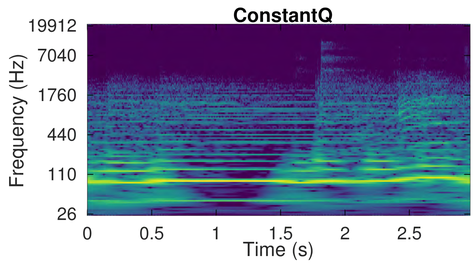
Prior work [1] showed that preserving phase information and sparse coding allowed for state-of-the-art vocal separation. **Will a "neurally-inspired" spectral representation (ConstantQ) show similar benefits over the commonly used Fourier transform?**

Fourier vs. ConstantQ

The Fourier and ConstantQ transforms take a signal and decompose it into its constituent frequencies, effectively allowing one to view the notes and chords played at any given time in a song.



- Linear spacing of frequency bands
- Poor resolution at relevant frequencies



- Logarithmic spacing of frequency bands (piano keys)
- Optimal resolution across spectrum

Results and Conclusions

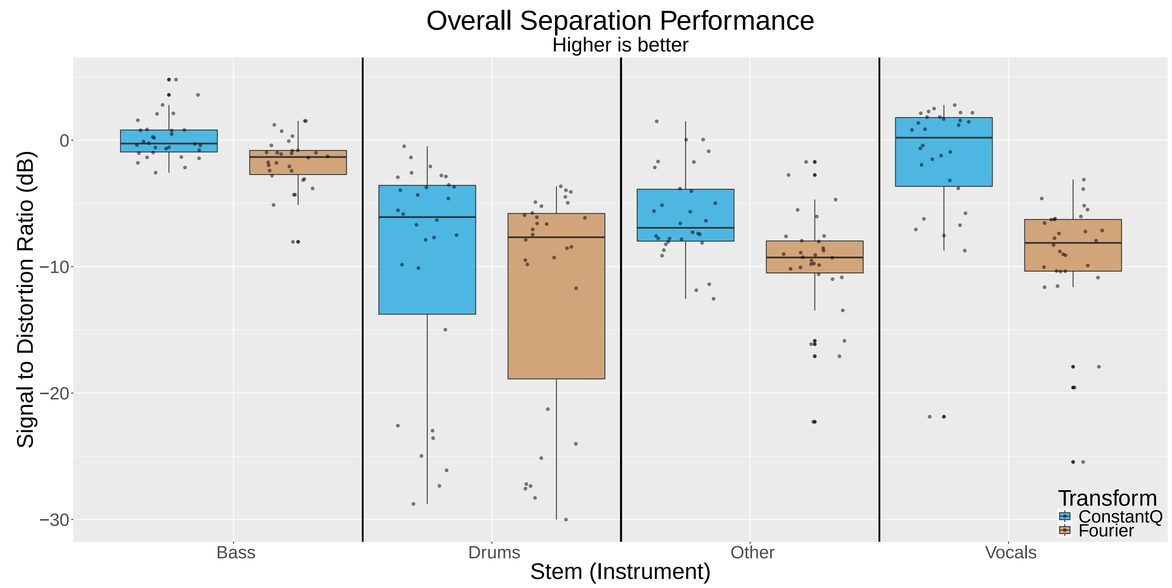


Fig. 1 Box-plot of overall performance (SDR) for ConstantQ (blue) and Fourier (brown) on testing set

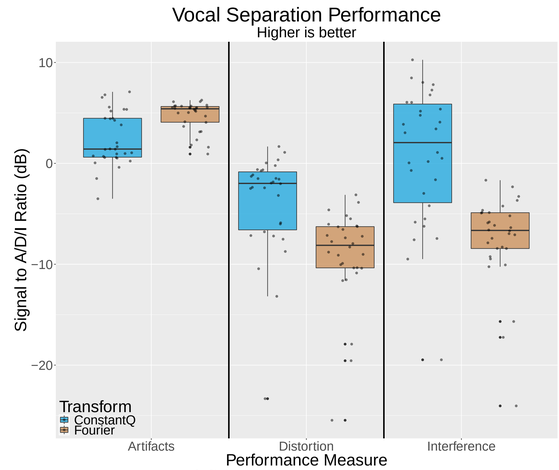


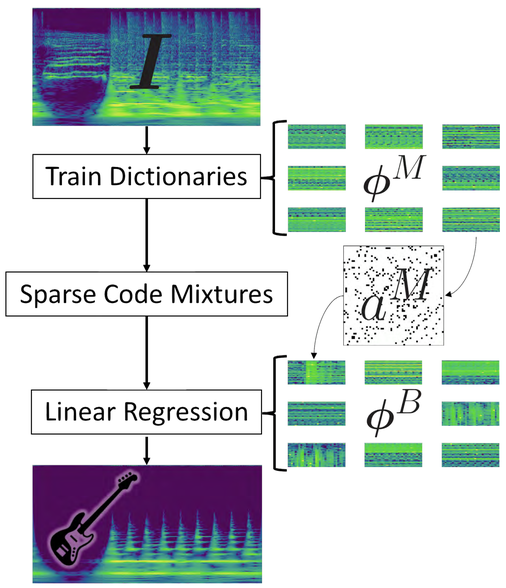
Fig. 2 Box-plot of SAR/SDR/SIR for vocals separation

Takeaways

1. **ConstantQ outperforms Fourier on overall separation (SDR) for all stems (Fig. 1)**
2. **While ConstantQ separates stems better (SIR), Fourier introduces slightly fewer artifacts (SAR) (Fig. 2)**

With a linear regression trained over 50 epochs, we separated bass, drums, vocals, and "other" stems from the sparse codes and reconstructed 10 minutes of audio using the phase of the original mixture. Using the BSS Eval Toolbox [2], we measured the amount of interference of other stems (SIR), artifacts introduced (SAR), and overall distortion (SDR) in our separated audio stems.

Methods



Training: Using spectral representations of mixtures of four individual stems as input, I, we use PetaVision to train dictionaries of convolutional features, $\boldsymbol{\phi}^M$. By minimizing the following energy function with respect to $\boldsymbol{\phi}$:

$$E = \frac{1}{2}||\overbrace{\vec{I} - \vec{\phi}\vec{a}}^{\text{Residual}}||_2^2 + \underbrace{\lambda||\vec{a}||_1}_{\text{Sparsity Penalty}}$$

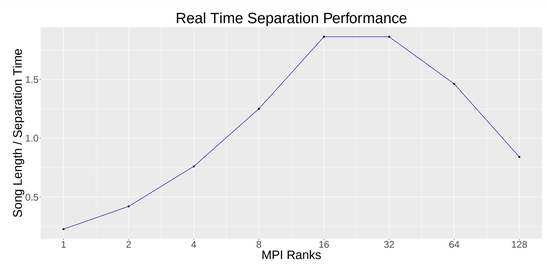
this technique finds an optimal sparse representation, a^M, of the input, and optimizes the features using gradient descent to represent generators of the input.

Sparse Code Mixtures: Once optimal dictionaries are trained, the sparse codes of the power spectra mixtures, a^M, are written out into training and testing sets.

Linear Regression: A linear regression is then trained by using the training set of sparse codes a^M_{train} to reconstruct power-spectra of the individual stems (i.e. Bass), resulting in new dictionaries of convolutional features (i.e. $\boldsymbol{\phi}^B$). These dictionaries are then used to separate out the stems from a^M_{test} and signal to distortion (SDR), interference (SIR), and artifact (SAR) ratios for the separated tracks are measured.

Versatile Scaling

Finally, we measured the computational performance of our classification model on a single full-length (three minute) song to investigate whether our model can separate stems in real-time. By slicing the song into N equal-sized pieces, we were able to distribute computation across multiple MPI ranks efficiently (see right). By distributing the computation across 16 or 32 MPI ranks, **our model was able to separate four individual stems faster than real time**.



References

[1] M. Dubey et al. Does phase matter for monaural source separation? In *arXiv '17*
[2] E. Vincent, R. Gribonval and C. FÃĺvotte Performance measurement in blind audio source separation In *IEEE Trans. Audio, Speech and Language Processing*

Acknowledgements

Support for this work was provided by U.S. Department of Energy at Los Alamos National Laboratory supported by Contract No. DE-AC52-06NA25396. We use PetaVision, a C++ library for designing and deploying large-scale neurally-inspired computational models. We would like to thank our mentor Garrett Kenyon as well as Austin Thresher, Nils Carlson, and Jacob Caroll for all their help with PetaVision. We would also like to thank Kris Garrett, Bob Robey, and Hai Ah Nam for giving their time to direct the PCSRI and provide students with this amazing opportunity. Big thank you to the Darwin admins for all the power nodes. LA-UR-18-27058

A Future Direction

In the future, we will investigate other neurally inspired approaches, including reintroducing phase information as well as modeling hemispheric lateralization in the brain to optimize our our separation results.